Sample of Knowledge Discovery Applications Capabilities at the University of Texas at Dallas

24 January 2006

Outline

- What is Knowledge Discovery?
- Prof. Bhavani Thuraisingham's Research
  - Text and Image Mining
    - Early research funded by MITRE, the Community Management Staff, Office of Research and Development (now AAT), National Imagery Mapping Agency (now NGA)
  - Suspicious Event Detection
  - Geospatial Data Integration and Mining
    - Partially funded by CH2MHILL
  - Assured Information Sharing
    - Air Force Office of Scientific Research
  - Biometrics (backup)

Outline -II

- Prof. Latifur Khan's Research
  - Multimedia/Image data extraction/Mining (Nokia)
    - PhD research at University of Southern California and now continuing at UTD
  - Intrusion detection
  - Web Page Prediction (NSF)
  - Bioinformatics (backup)
- Prof. Murat Kantarcioglu Research
  - Privacy/Security Preserving Data Mining
    - PhD research at Purdue U; and now continuing at UTD
  - Misinformation / Insider Threat
    - White paper being prepared for AFOSR

Outline - III

- Prof. Kang Zhang’s Research (Backup)
  - Knowledge Discovery and Visualization (NSF)
- Our Vision for Research
  - Assured Information Sharing and Knowledge Discovery
- Our Current Collaborations
- Some Past Efforts for Federal Government

What is Knowledge Discovery (KDD)?

Information Harvesting

Data Mining

Knowledge Mining

Knowledge Discovery in Databases

Data Dredging

Data Archaeology

Data Pattern Processing

Database Mining

Knowledge Extraction

Siftware

The process of discovering meaningful new correlations, patterns, and trends by sifting through large amounts of data, often previously unknown, using pattern data Mining: Technologies, Techniques, Tools and Trends, CRC Press, Thuraisingham 1998)

Knowledge Discovery in Text

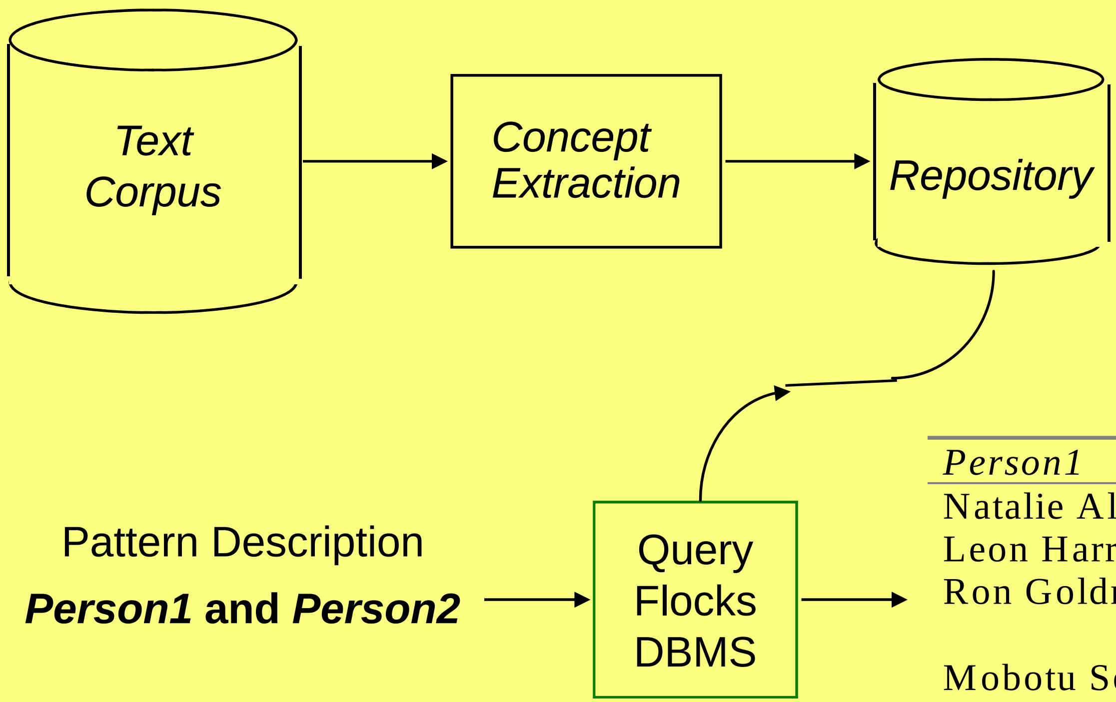
Text Corpus → Concept Extraction → Repository → Association Rule Product

Goal: Find Cooperating/ Combating Leaders in a territory

Person1	*Person2*	
Natalie Allen	Linden Soles	117
Leon Harris	Joie Chen	53
Ron Goldman	Nicole Simpson	19
	. . .	
Mobotu Sese Seko	Laurent Kabila	10

Too Many Results

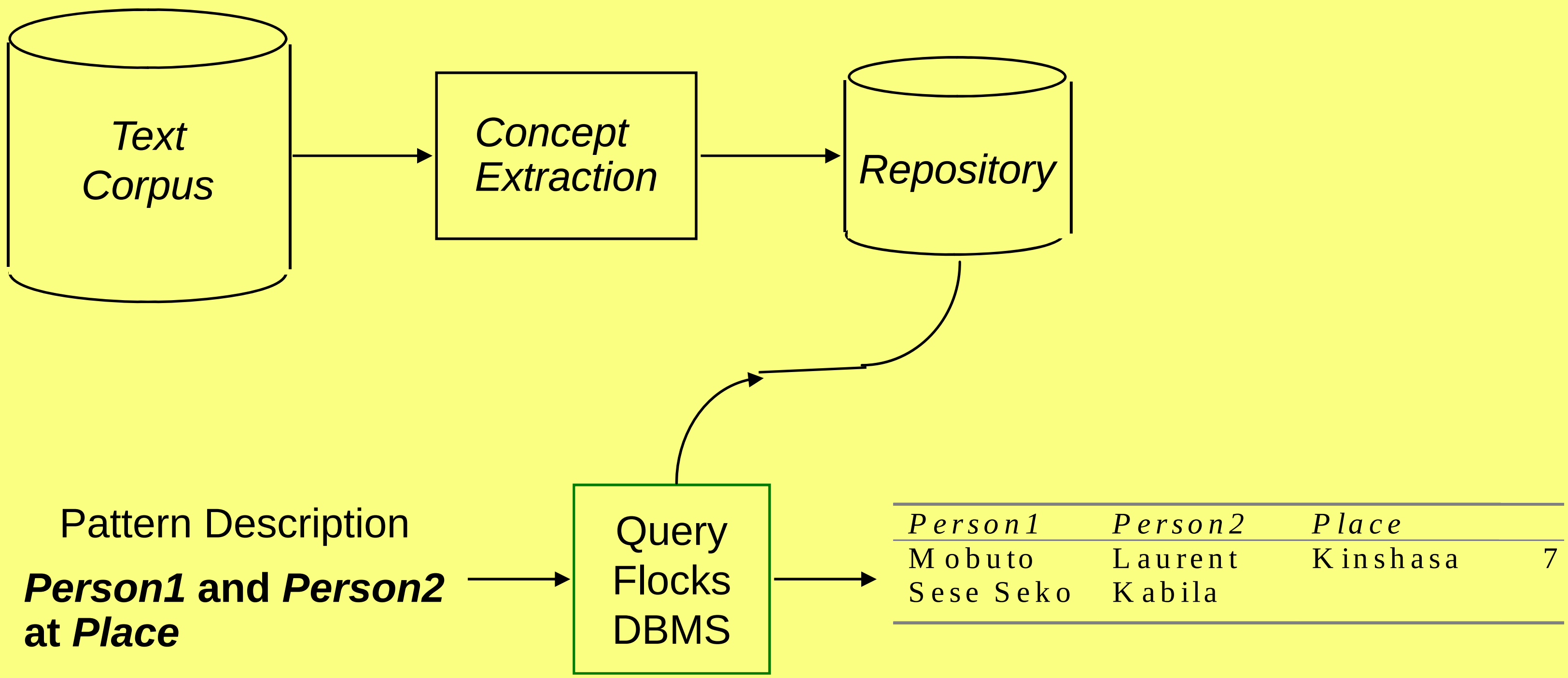
Query flocks Tool



Person1	*Person2*	
Natalie Allen	Linden Soles	117
Leon Harris	Joie Chen	53
Ron Goldman	Nicole Simpson	19
	. . .	
Mobotu Sese Seko	Laurent Kabila	10

Still Too Many Results

Query Capability



Person1	*Person2*	*Place*	
Mobuto	Laurent	Kinshasa	7
Sese Seko	Kabila		

Knowledge Discovery in Images

- **Goal: Find *unusual* changes**
  Process:
  - **Use data mining to model normal differences between images**
  - **Find places where differences don't match model**
- **Questions to be answered:**
  - **What are the right mining techniques?**
  - **Can we get useful results?**

Change Detection:

- Trained Neural Network to predict “new” pixel from “old” pixel
  - Neural Networks good for multidimensional continuous data
  - Multiple nets gives range of “expected values”
- Identified pixels where actual value substantially outside range of expected values
  - Anomaly if three or more bands (of seven) out of range
- Identified groups of anomalous pixels

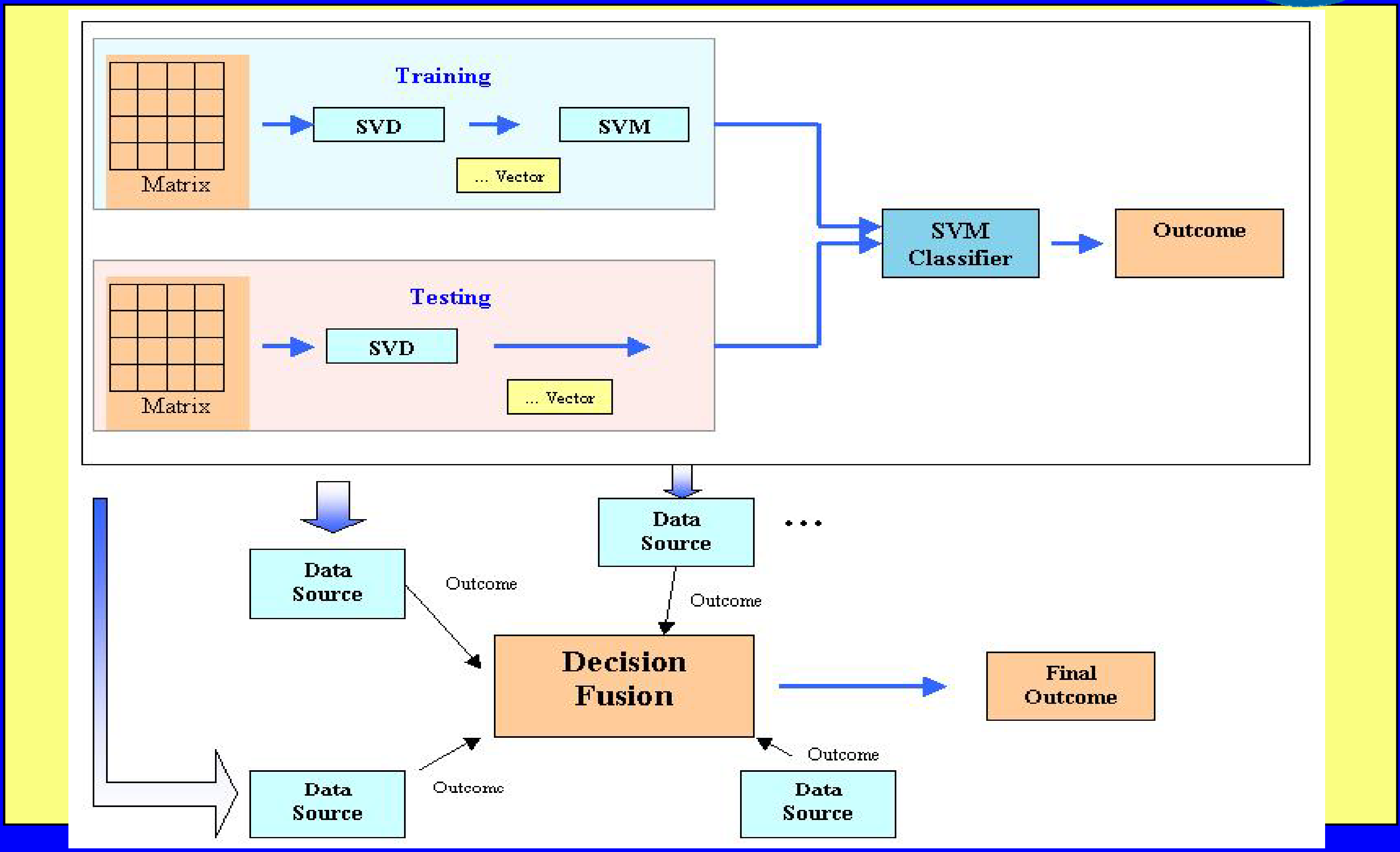
Data Mining for Suspicious Event Detection

- **We define an event representation measure based on low-level features**
- **Having a well-defined event representation allows us to compare events. Our desired effect is that video events that contain the same semantic content will have small dissimilarity from one another (i.e. be perceived as the same event).**
- **This allows us to define “normal” and “suspicious” behavior and classify events in unlabeled video sequences appropriately**
- **A visualization tool can then be used to enable more efficient browsing of video data**

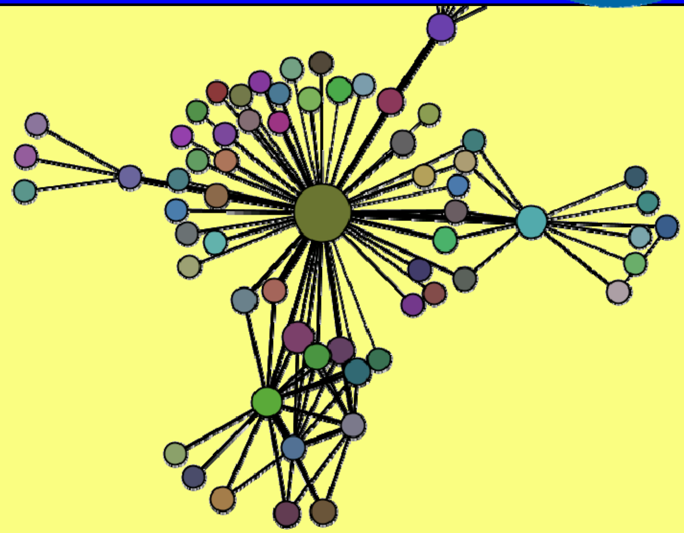
Data Mining for Fraudulent Claims Detection

- **Work for the State of Texas; Inspector General of Texas**
- **Purchased a 16 Terabyte Sun Server**
- **Oracle database management**
- **Claims Data of about 11 terabytes from the state**
- **Ensuring Privacy by removing elements that can reveal identity**
- **Data Mining to determine fraudulent claims**
- **Also implementing Privacy constraint processing techniques for ensuring privacy**
- **Plan to show demonstration also to Pharmaceutical companies as permitted by the State of Texas**

Geospatial Data Integration

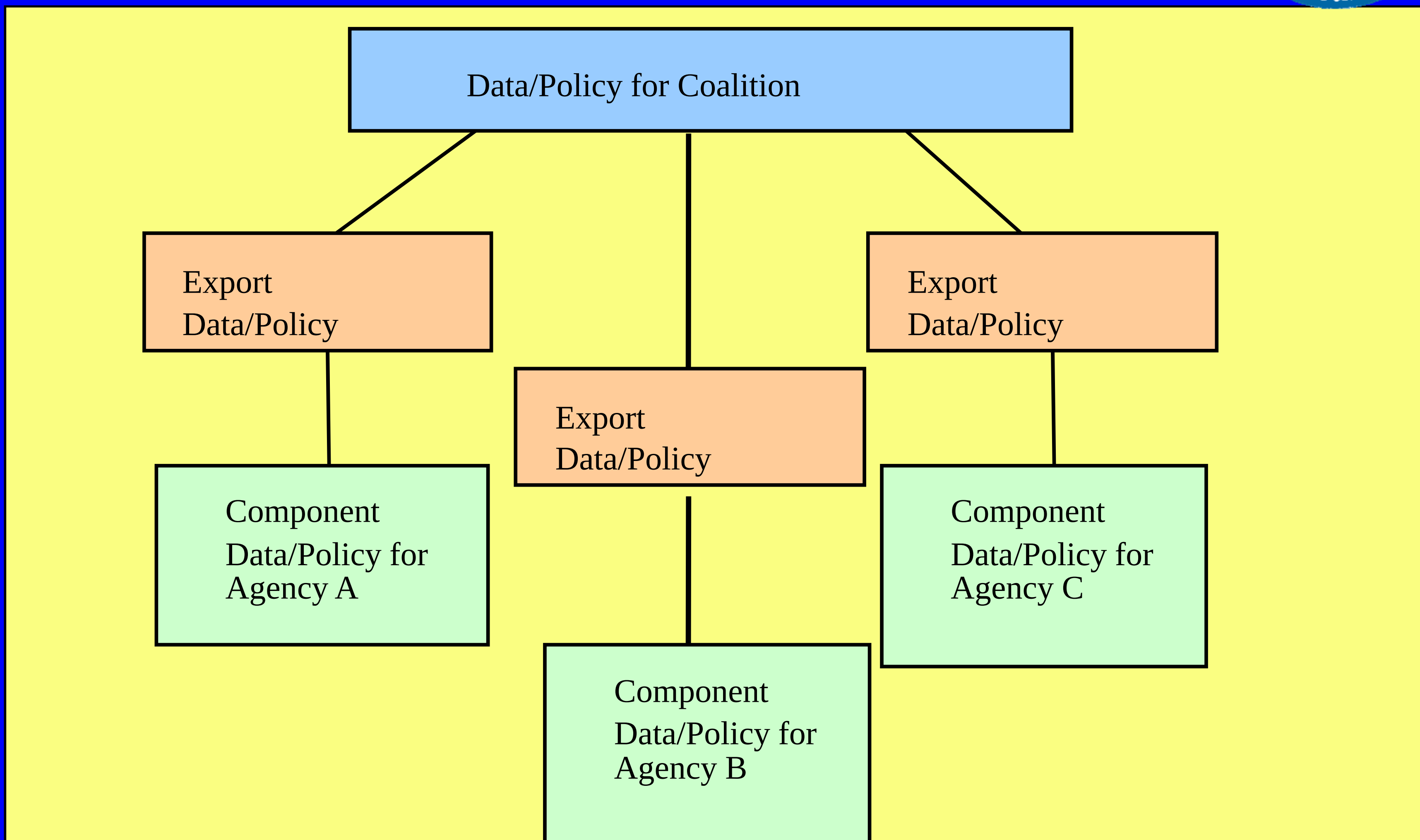


Social Network Analysis



- Suspicious Message Detection
  - Adaptation of existing spam detection techniques
    - Naïve Bayesian Classification
    - Support Vector Machines
    - Keyword Identification
- Application of graph theory on existing social network techniques
  - Detection of roles
  - Detecting individuals that stray outside known social circles
- Detecting chains of conversation through message correlation analysis
  - Determination of word frequencies within a message
  - Comparison between existing suspicious messages
  - Adaptive scoring system that uses the intersection of word content to determine how strongly messages or conversations correlate

Assured Information Sharing Across Coalitions

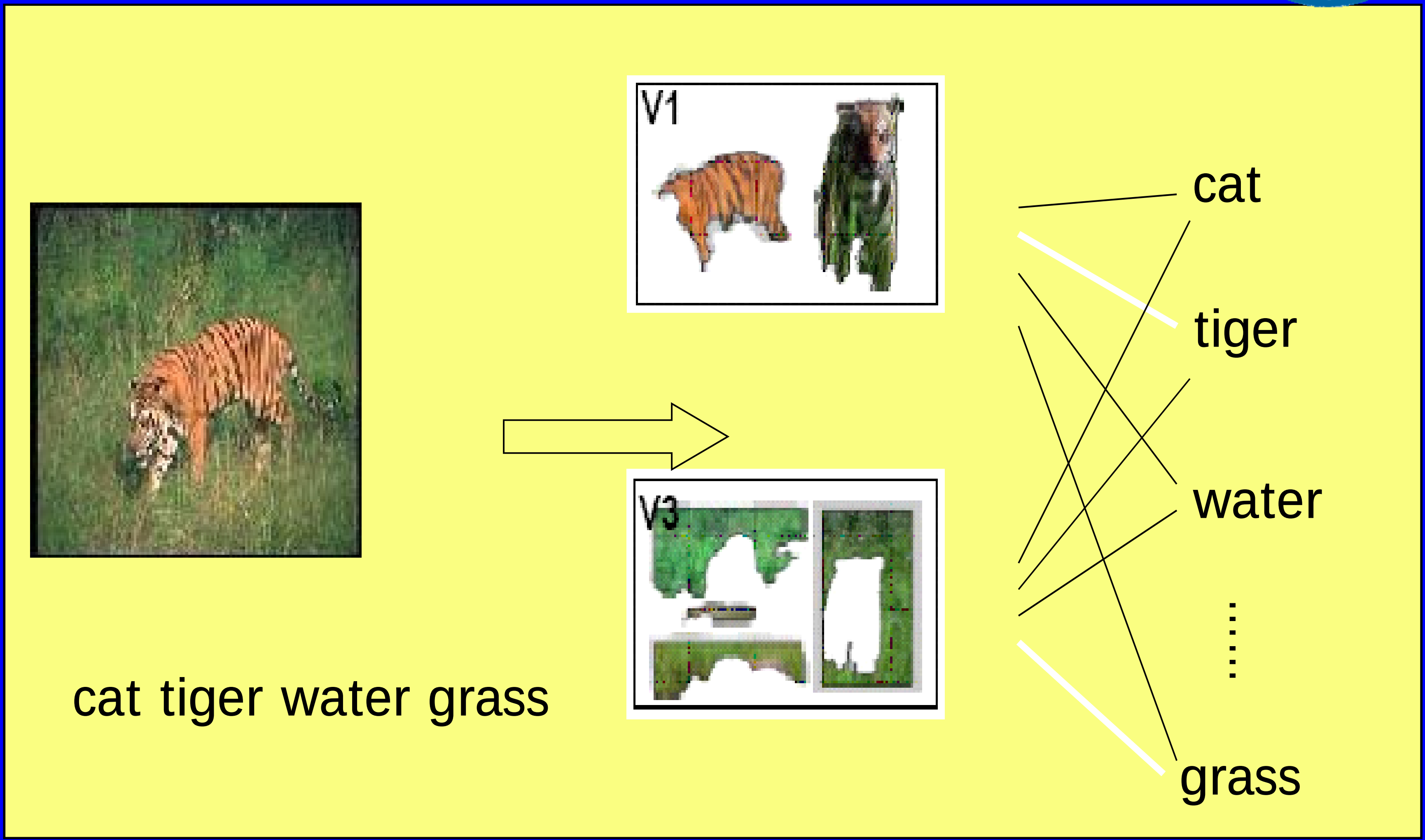


Multimedia/Image Mining

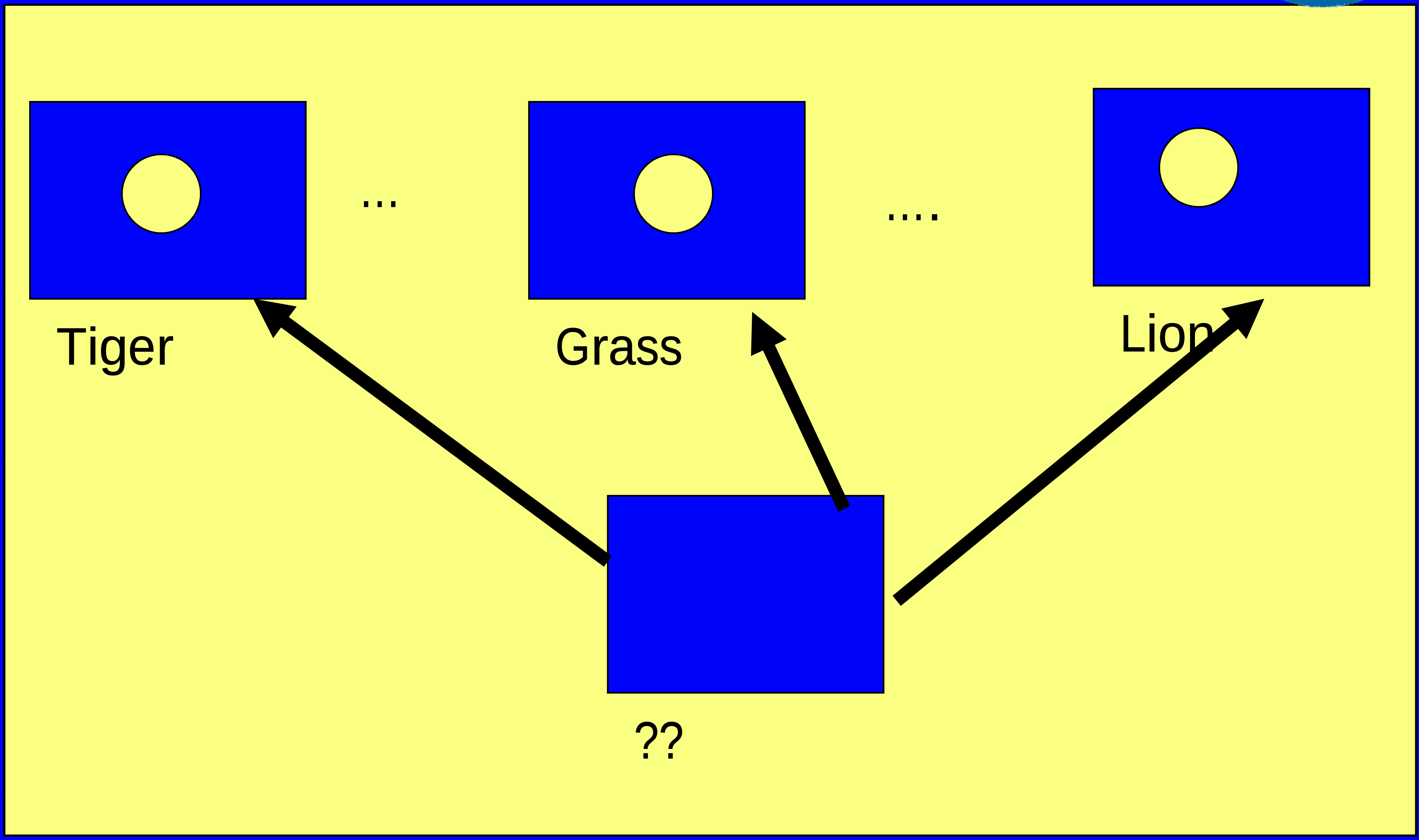
Automatically annotate images then retrieve based on the textual annotations.

Images **Segments** **Blob-tokens**

Multimedia/Image Mining: Correlation

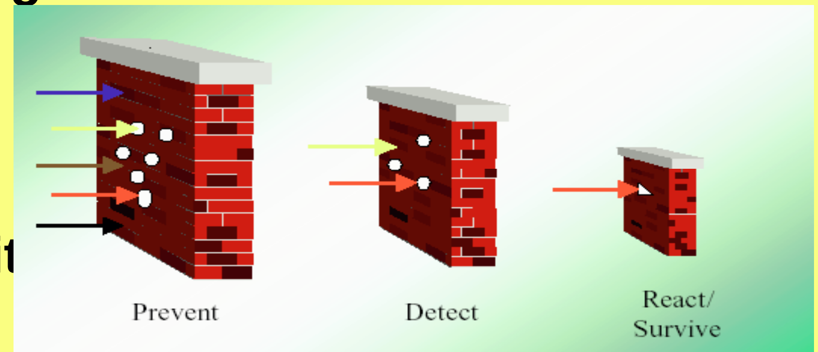


Multimedia/Image Mining: Auto Annotation



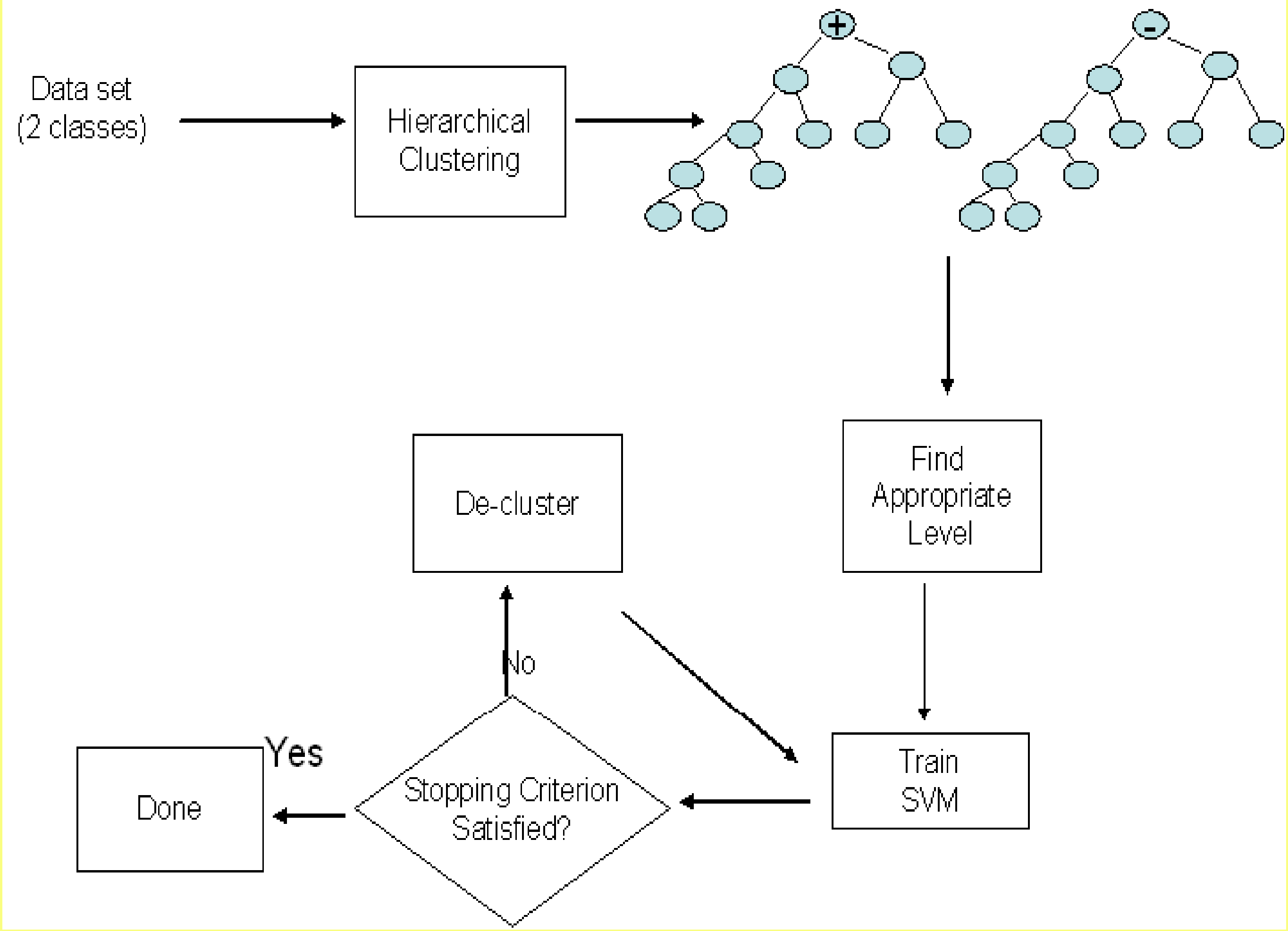
Intrusion Detection

- An intrusion can be defined as "any set of actions that attempt to compromise the integrity, confidentiality, or availability of a resource".
- Intrusion detection systems are split into two groups:
  - Anomaly detection systems
  - Misuse detection systems
- Use audit logs
  - Capture *all* activities in network and hosts.
  - But the amount of data is huge!
- Goal of Intrusion Detection Systems (IDS):
  - To detect an intrusion as it happens and be able to respond to it
  - Lower false positive
  - Lower false negative



Security principles: layered mechanisms

Intrusion Detection: Solution



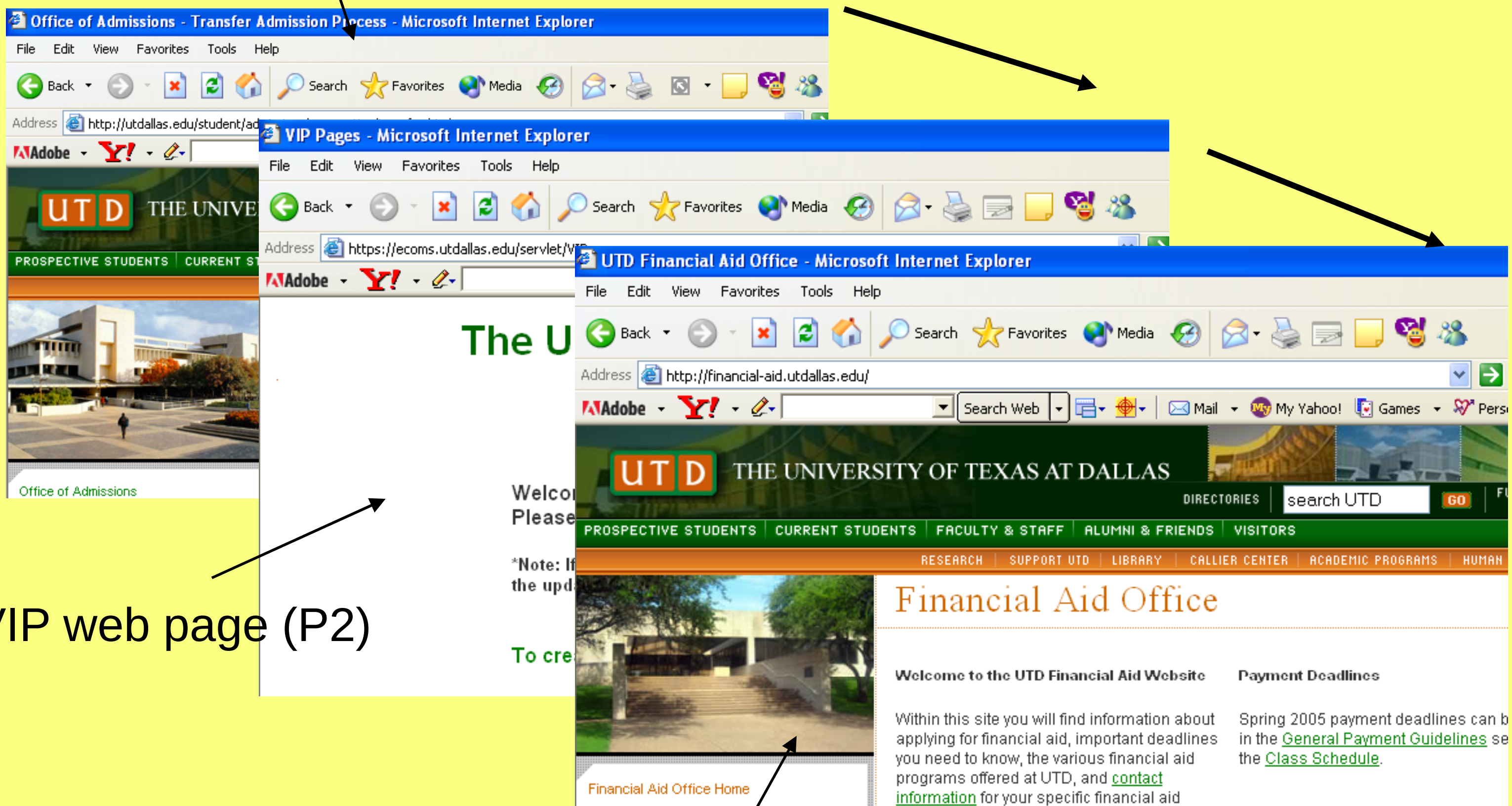
Intrusion Detection: Results

Training Time, FP and FN Rates of Various Methods

Methods	Average Accuracy	Total Training Time	Average FP Rate (%)	Average FN Rate (%)
Random Selection	52%	0.44 hours	40	47
Pure SVM	57.6%	17.34 hours	35.5	42
SVM+Rocchio Bundling	51.6%	26.7 hours	44.2	48
SVM + DGSOT	69.8%	13.18 hours	37.8	29.8

Web Page Prediction: Problem Description

Office of admission (P1)



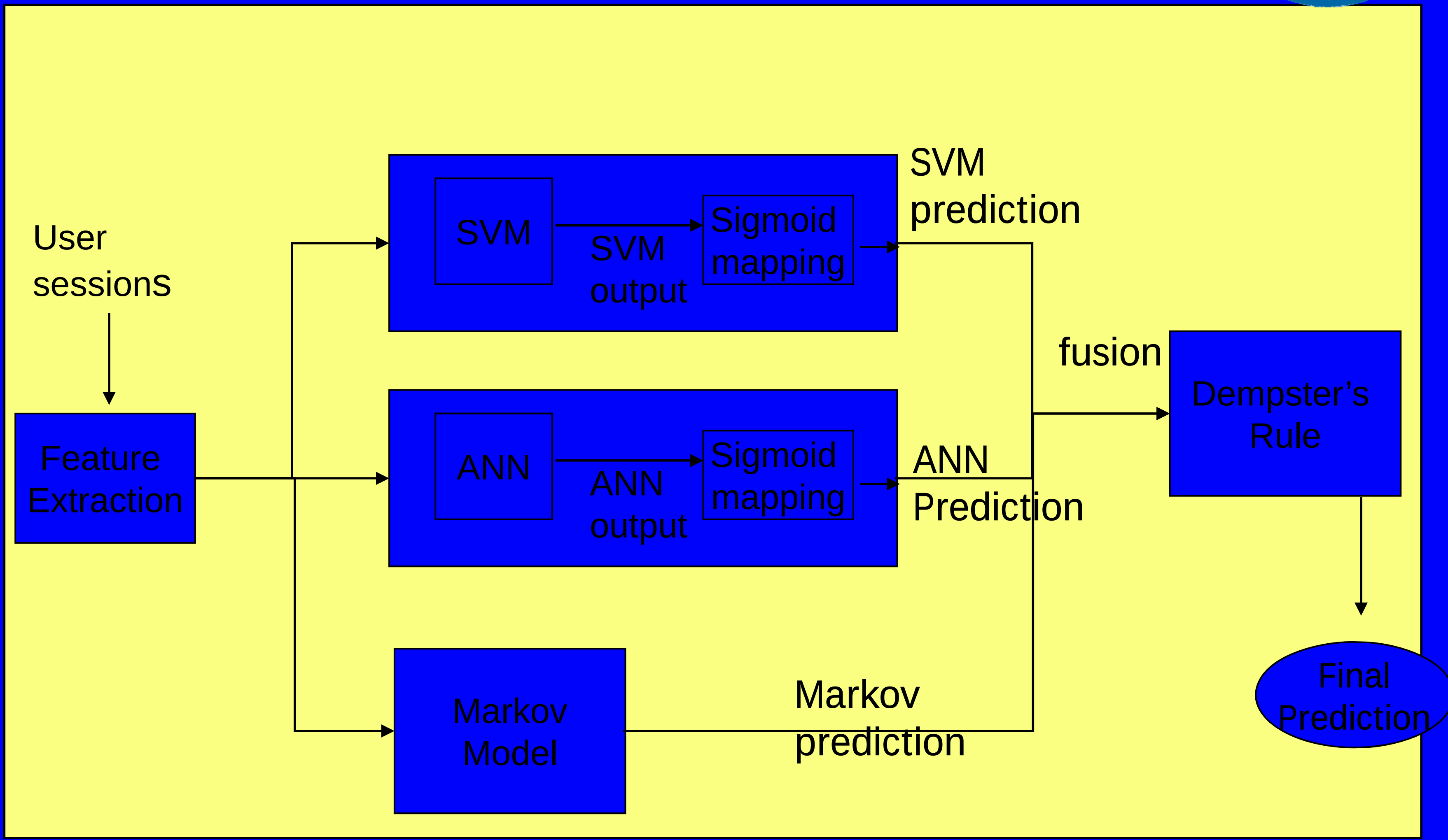
?

What page is Next??

VIP web page (P2)

Financial Aid Information (P3)

Web Page Prediction: Architecture



Web Page Prediction: Feature Extraction

- **Sliding Window**

A <**1 ,2, 3**, 4, 5, 6>

A <1 ,**2, 3, 4**, 5, 6>

A <1 ,2, **3, 4, 5**, 6>

A <1 ,2, 3, **4, 5, 6**>

Web Page Prediction: Results/one hop-rank 4

Table. : Using all probability measurements with one hop and rank 4.

Method	pr (match)	pr (hit|match)	pr (hit)	pr (miss | match)	pr(miss)	pr(hit)/ pr(miss)	overall pr(hit) /pr(miss)	pr(hit | mis-match)	overall accu-racy
ARM	0.592	0.211	0.125	0.788	0.467	0.268	0.143	0	0.125
SVM	0.592	0.298	0.177	0.701	0.415	0.425	0.315	0.154	0.24
ANN	0.592	0.308	0.182	0.691	0.409	0.445	0.332	0.164	0.249
Markov	0.592	0.35	0.207	0.64	0.385	0.539	0.262	0	0.207
ANN and Markov	0.592	0.346	0.205	0.653	0.387	0.53	0.368	0.157	0.269
SVM and Markov	0.592	0.351	0.208	0.648	0.384	0.542	0.375	0.158	0.273
SVM and ANN	0.592	0.297	0.176	0.702	0.416	0.423	0.314	0.154	0.239
SVM, Markov, ANN	0.592	0.348	0.206	0.651	0.386	0.534	0.368	0.154	0.269

Training accuracy

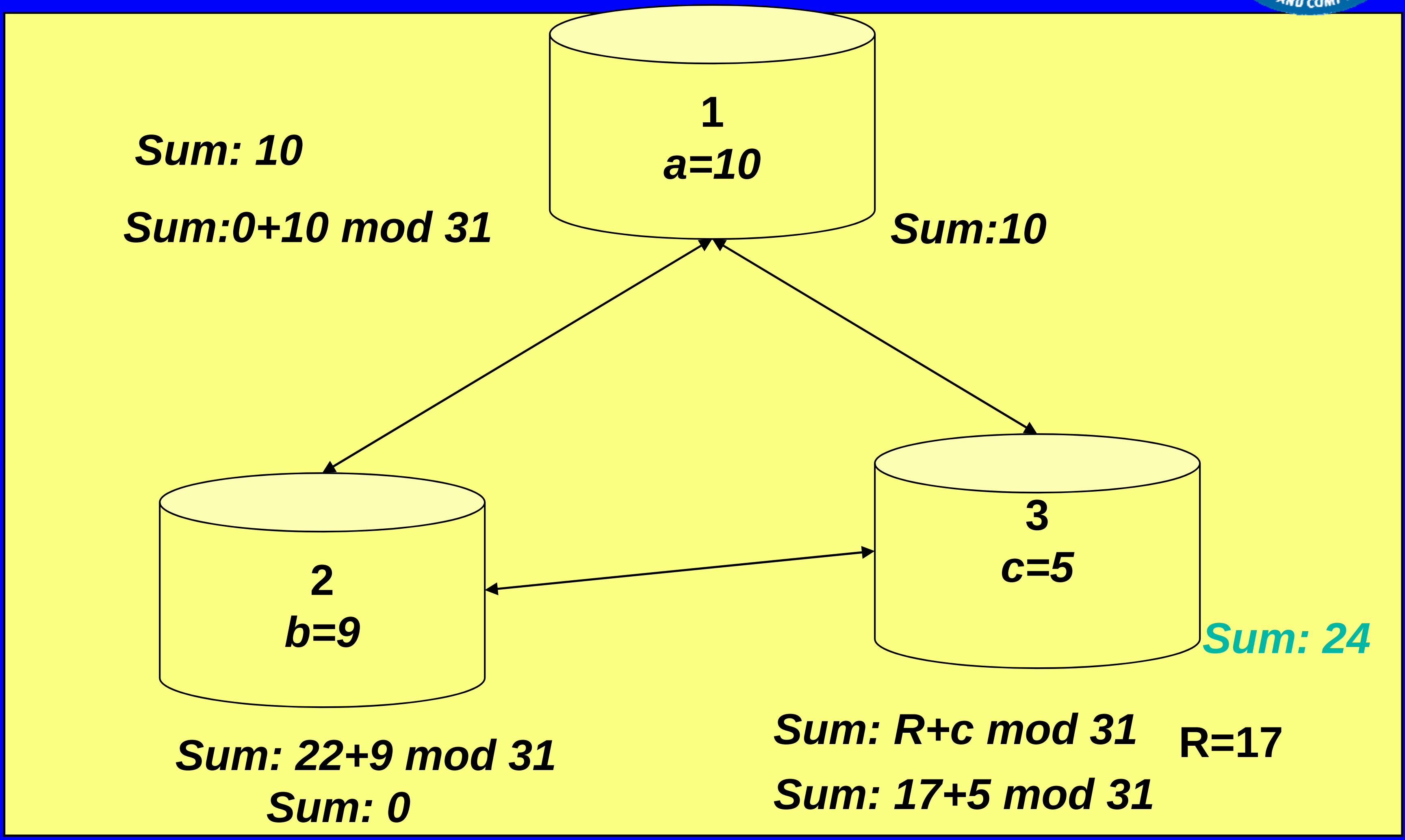
Generalization accuracy

overall accuracy

Privacy and Security Preserving Data Mining

- Goal of data mining is summary results
  - Association rules
  - Classifiers
  - Clusters
- The results alone need not violate privacy
  - Contain no individually identifiable values
  - Reflect overall results, not individual organizations
- Privacy-Preserving Distributed Data Mining: Why ?
  - Data needed for data mining maybe distributed among parties (Credit card fraud data, intelligence agency data)
- Inability to share data due to security or legal reasons
- Even partial results may need to be kept private

Securely Computing Summation



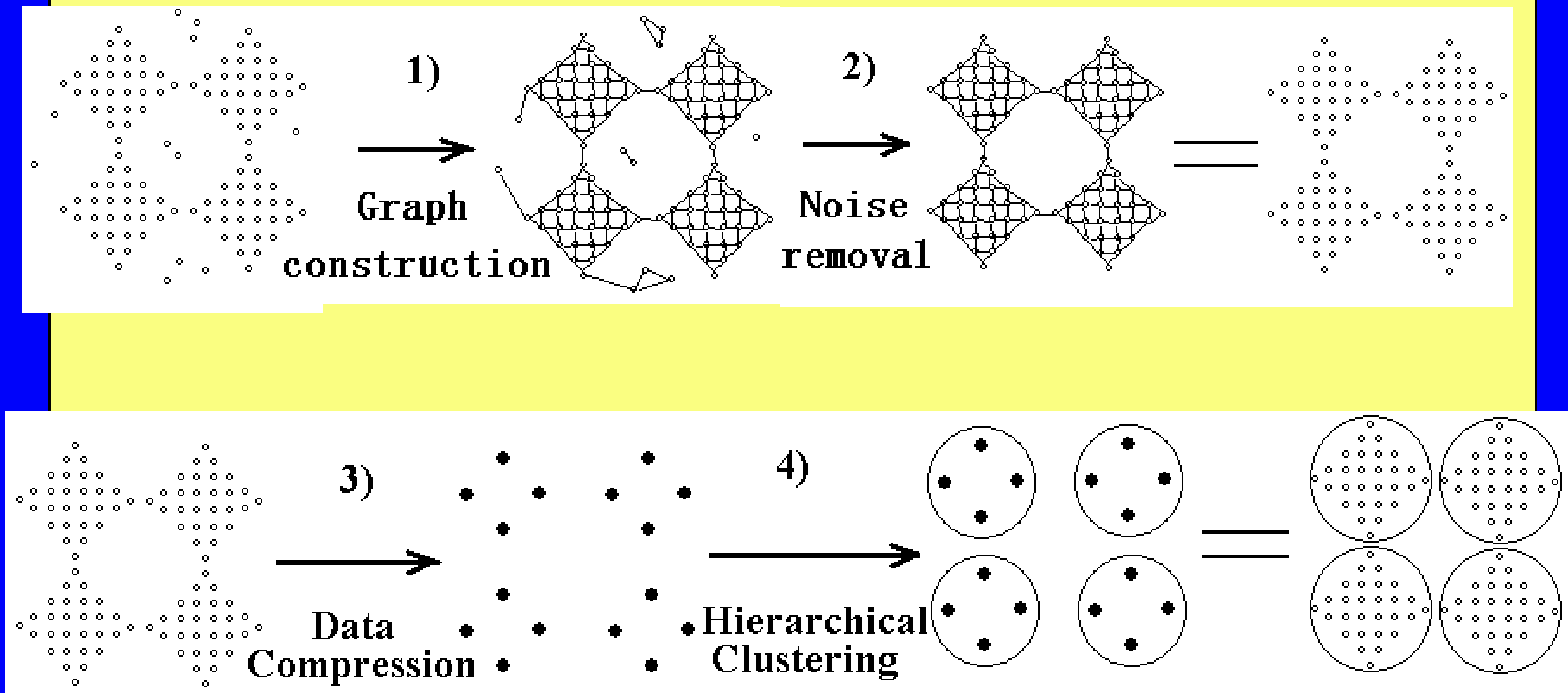
Tools Developed for Privacy Preserving Data Mining

- **Privacy-preserving Distributed Data Mining (PPDDM) Tools**
  - **Privacy-preserving association rule mining (TKDE ‘04, DMKD02)**
  - **Privacy-preserving k-NN classification (PKDD ‘04)**
  - **Privacy-preserving Naïve Bayes Classifier (ICDM, PSDM ’03)**
  - **Architecture for privacy-preserving data mining (ICDM, PSDM ’02)**
- **Secure toolbox for PPDDM (PKDD PSDM ‘04)**
  - **Common secure protocols used in PPDDM**
- **Using Data Mining results privately**
  - **Private Classification (DMKD ’03)**
  - **Privacy Implications of Data Mining Results (SIGKDD ’04)**

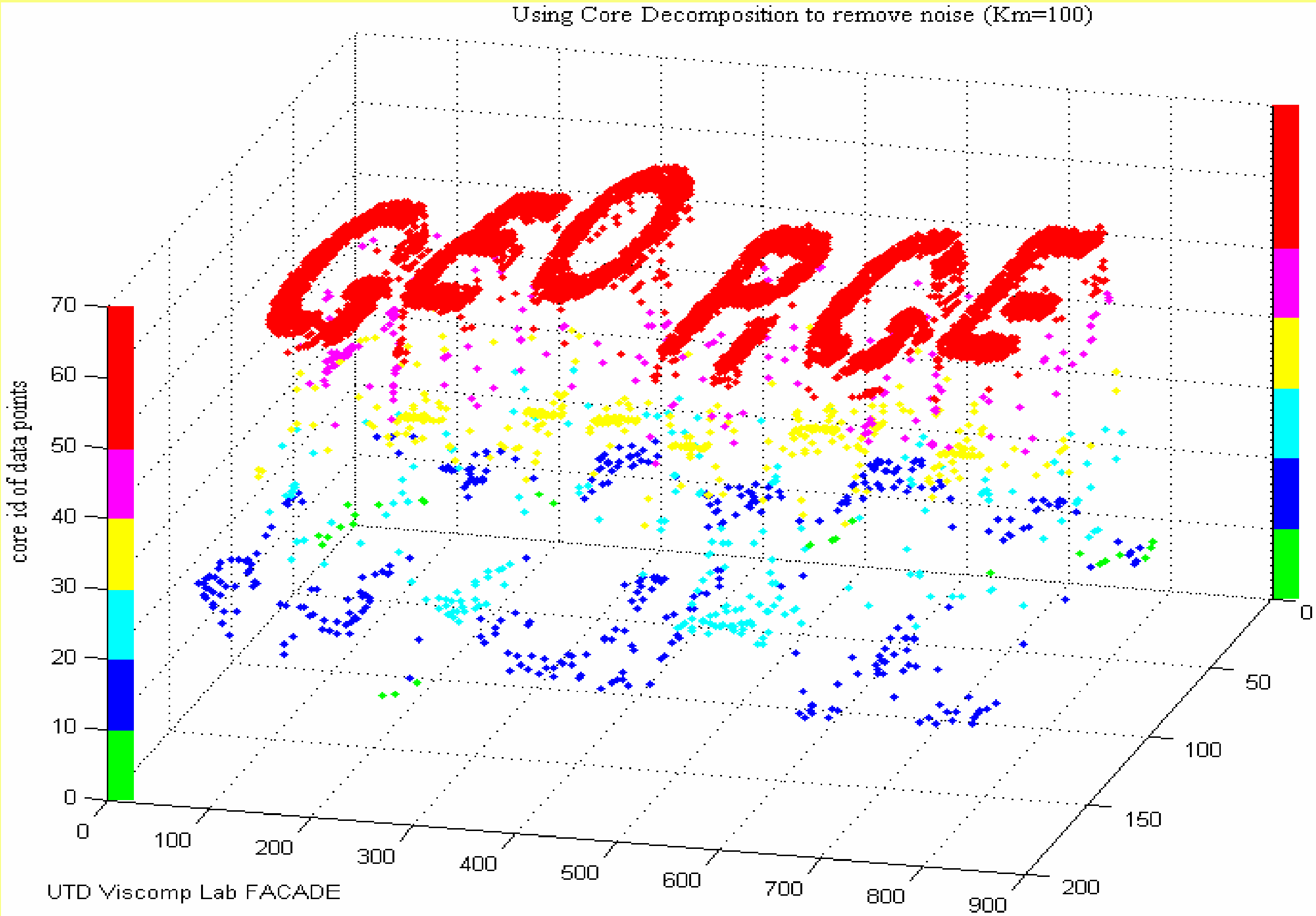
Misuse/Misinformation/ Insider threat (Murat)

- %50 of corporate breaches or losses of information that were made public in the past year were insider attacks
- %50 of those insider attacks were the thefts of information by employees
- It is hard to model individuals!!!
- Role based access control provides tools to model given roles
- Challenge: How to develop models for predicting normal usage of a role vs misuse?
- Challenge: How to integrate misuse, auditing and access control systems?
- Current Status: We are developing misuse detection system based on clustering.

FACADE (Fast and Automatic Clustering Approach to Data is featured:



Visualized Noise Removal



The core hierarchy

Some Experiences with Tools

- **Tools developed in-house**
  - **Query flocks, Image mining tool**
  - **Intrusion detection tool, Web page prediction tool**
  - **Multimedia mining/Image extraction including MPEG7 feature descriptors**
  - **Cluster visualization tool**
- **External tools**
  - **Oracle data mining product**
  - **IDIS data mining tool**
  - **WEKA data mining tool**
  - **Lockheed Martin's RECON**
  - **XML SPIE and QUIP**
  - **INTEL OpenCV**

Our Vision for Assured Information Sharing/KDD

Secure Grid *(NSA Unsolicited proposal)*

Game Theory *(AFOSR - current)*

Dependable Information Management *(NSF, ONR)*

Link Analysis *(DHS Proposal Requested, CIA Interest)*

Assured Information Sharing/KDD

Misinformation/ Misuse *(AFOSR white paper requested)*

Privacy Preserving data mining *(NSF)*

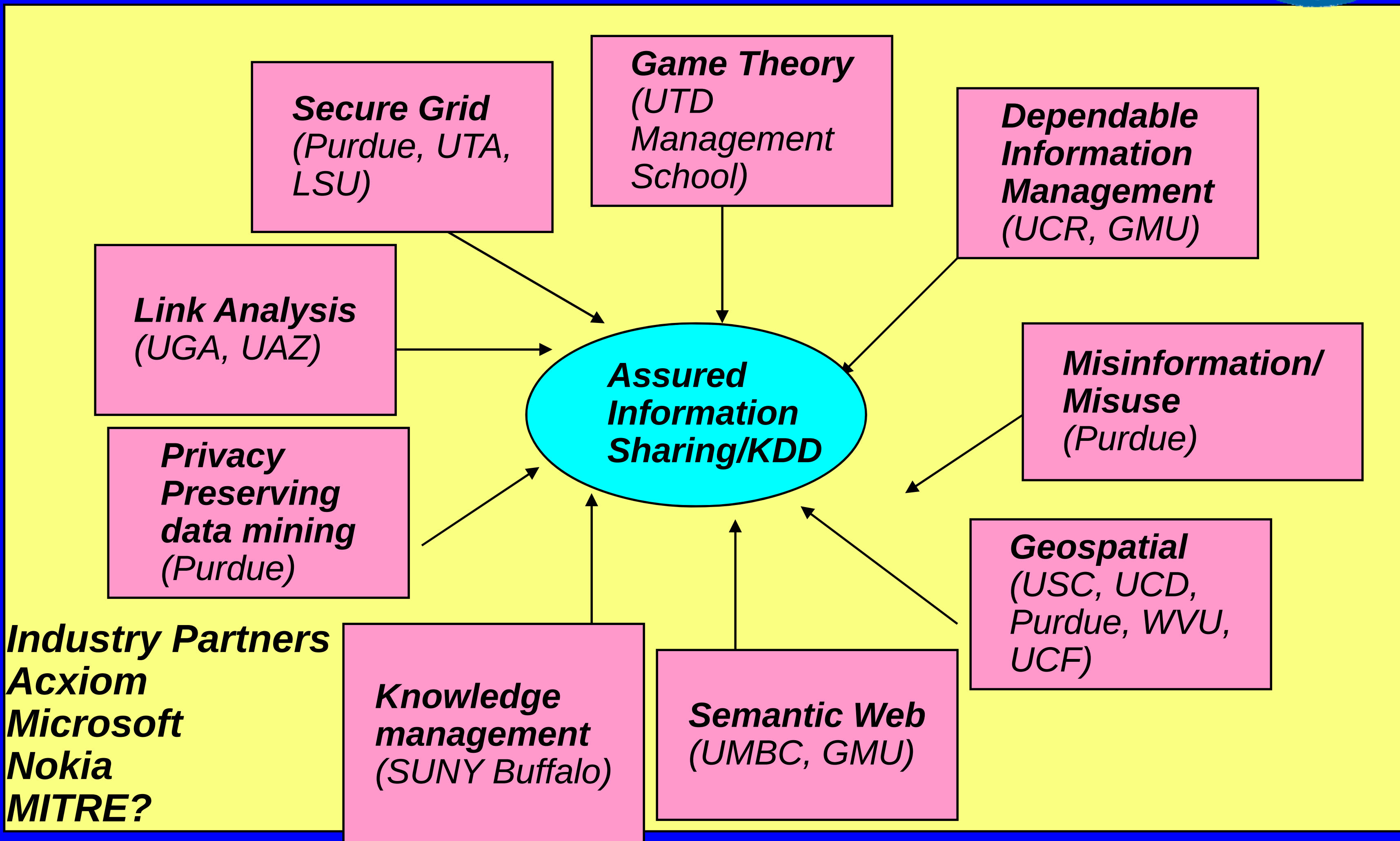
Geospatial *(NGA FY05 proposal selectable)*

Knowledge management *(AFRL Presentation requested)*

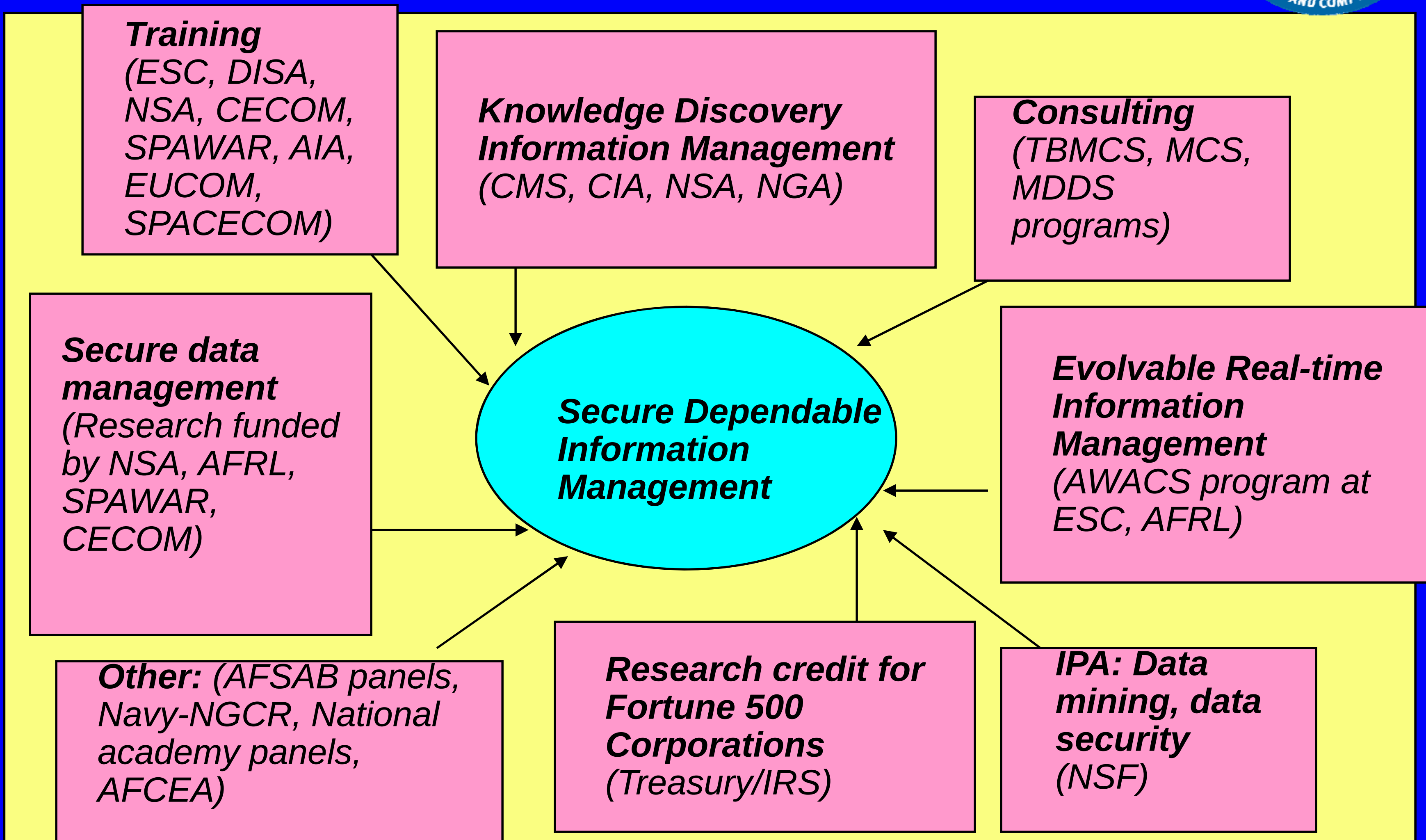
Semantic Web *(NSF, MURI ideas)*

Technologies will contribute to Assured Information Sharing

Our Collaborations in Assured Information Sharing and KDD

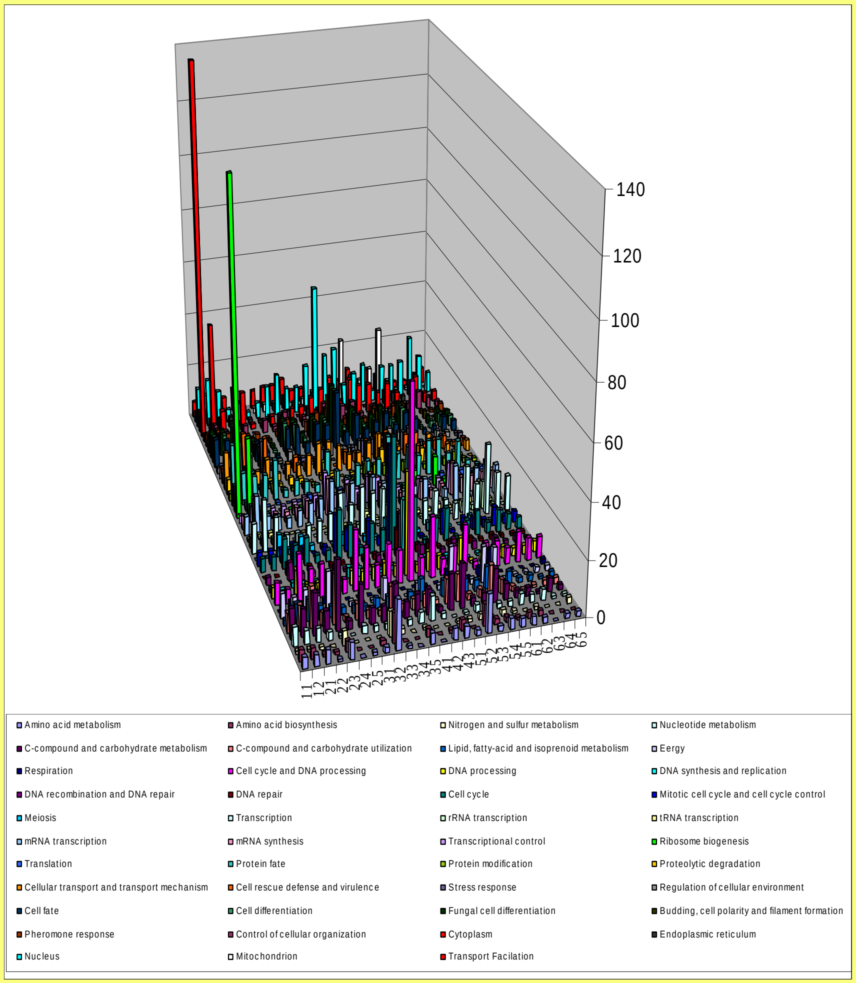


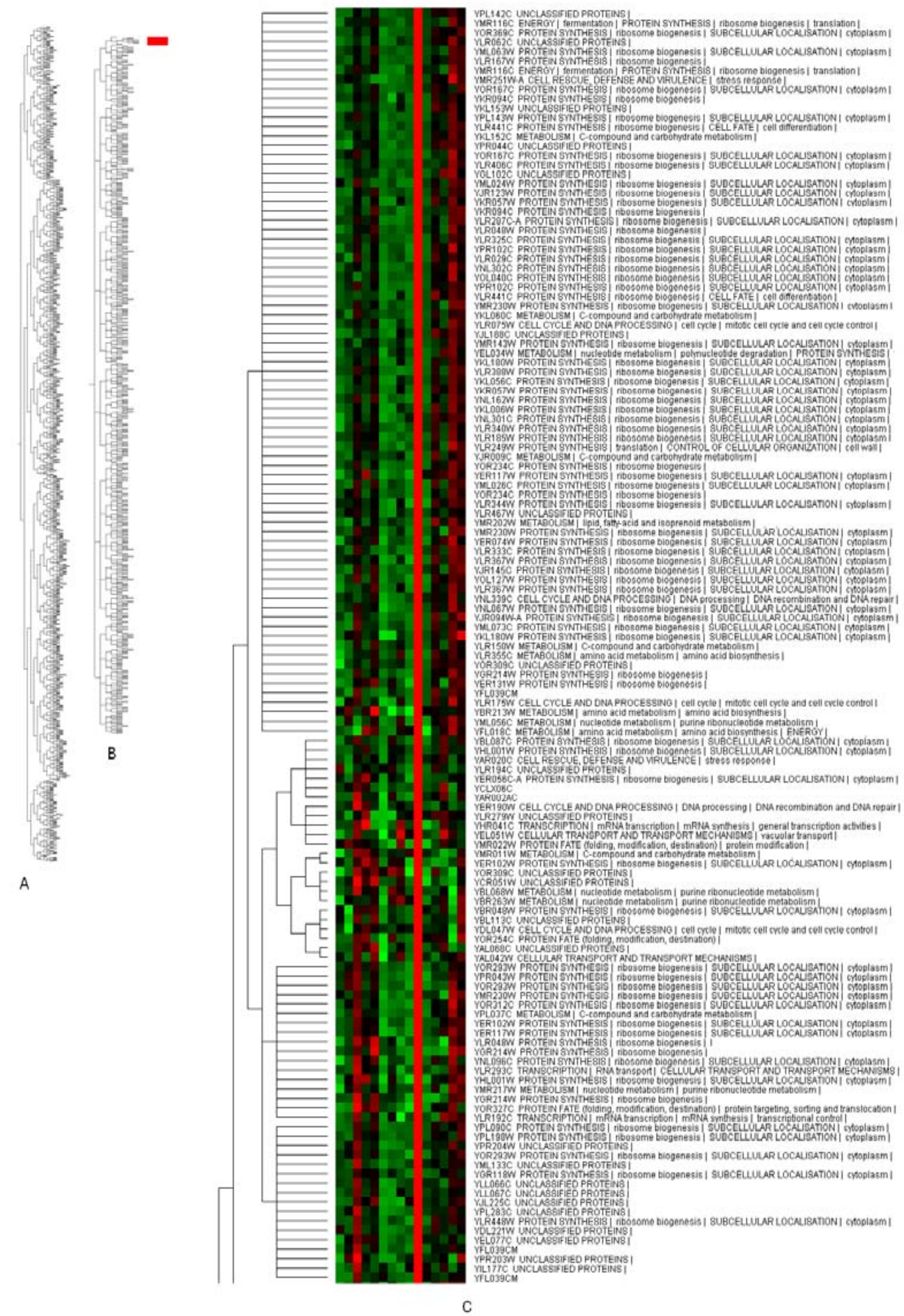
Some Previous Efforts for Federal Government



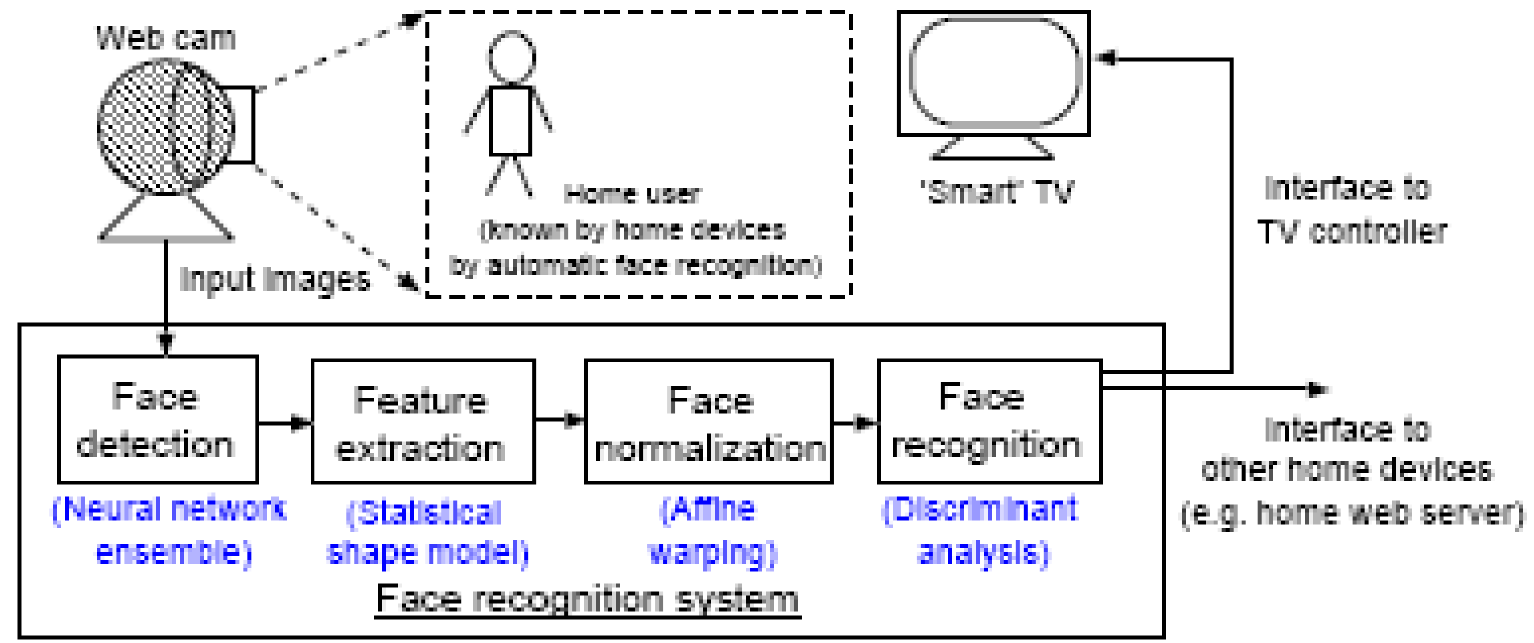
Backup Charts

Bioinformatics: Clustering Microarray Data

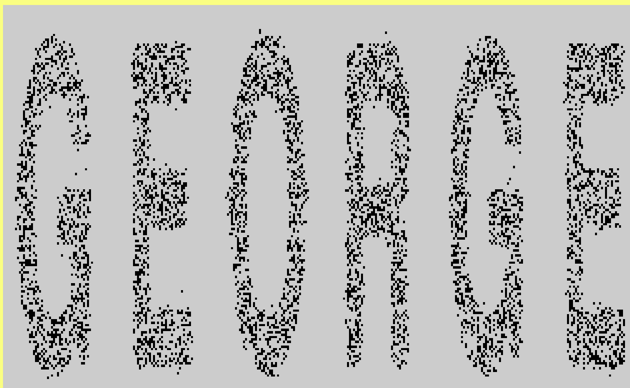




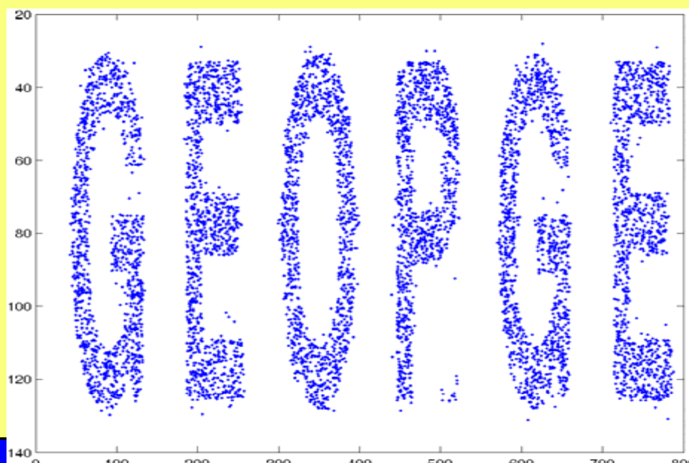
Biometrics: Face Recognition



Visualization: Customization

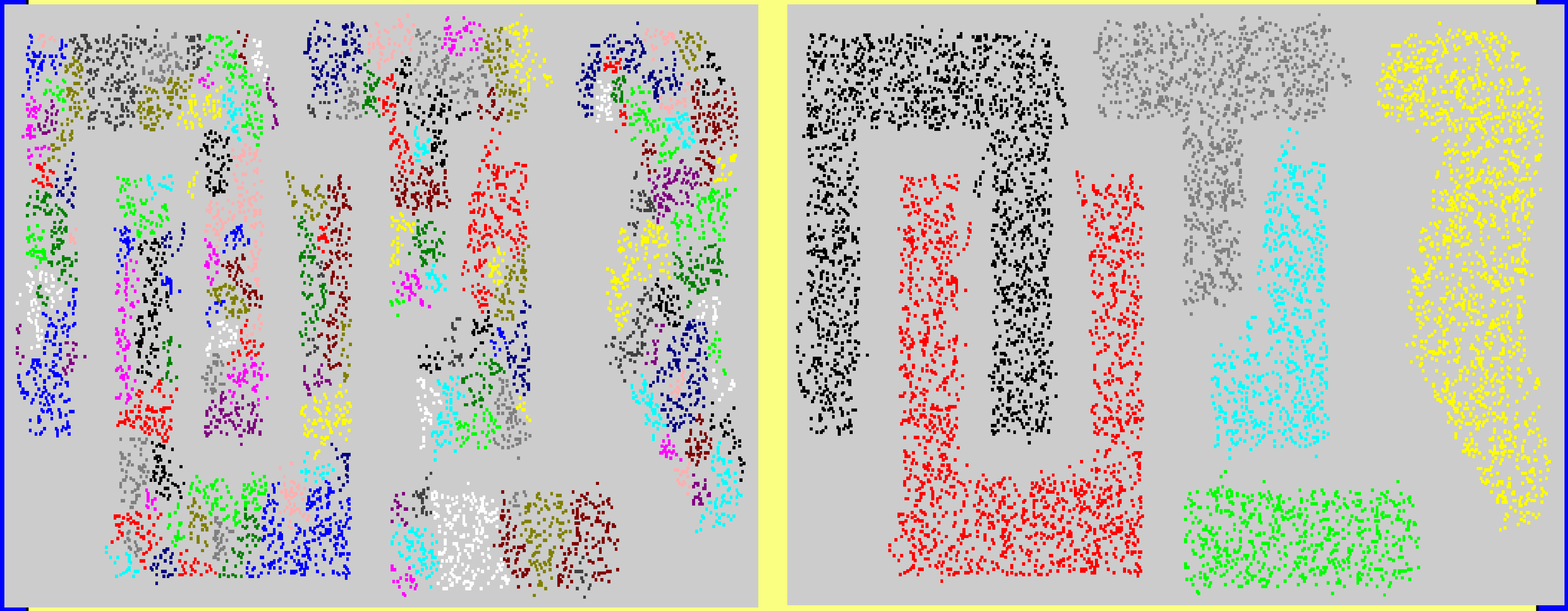


(coreID>45)

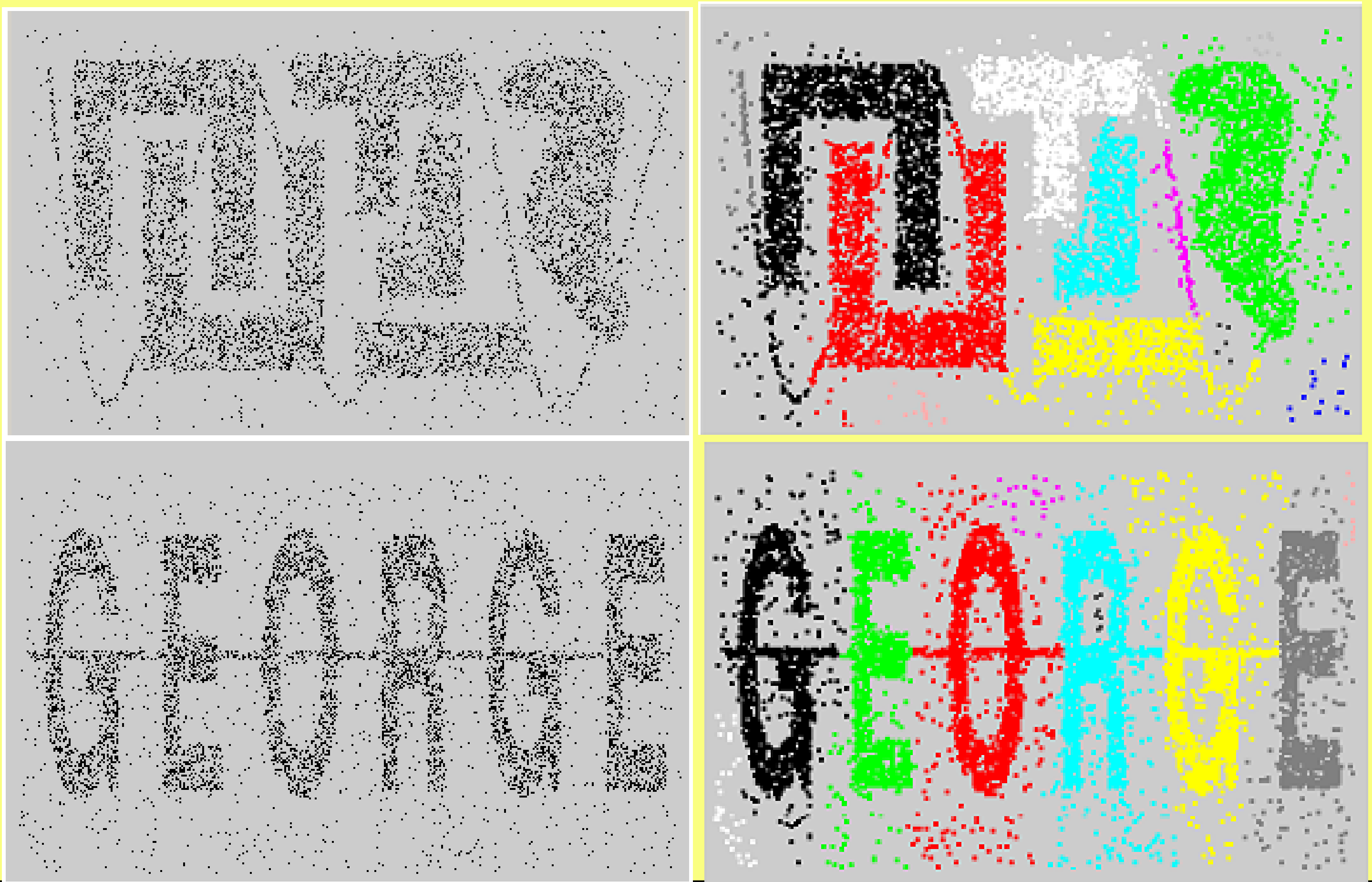


(((coreID>52) and (coreID<70))
or
((coreID>45) and (coreID<52)))

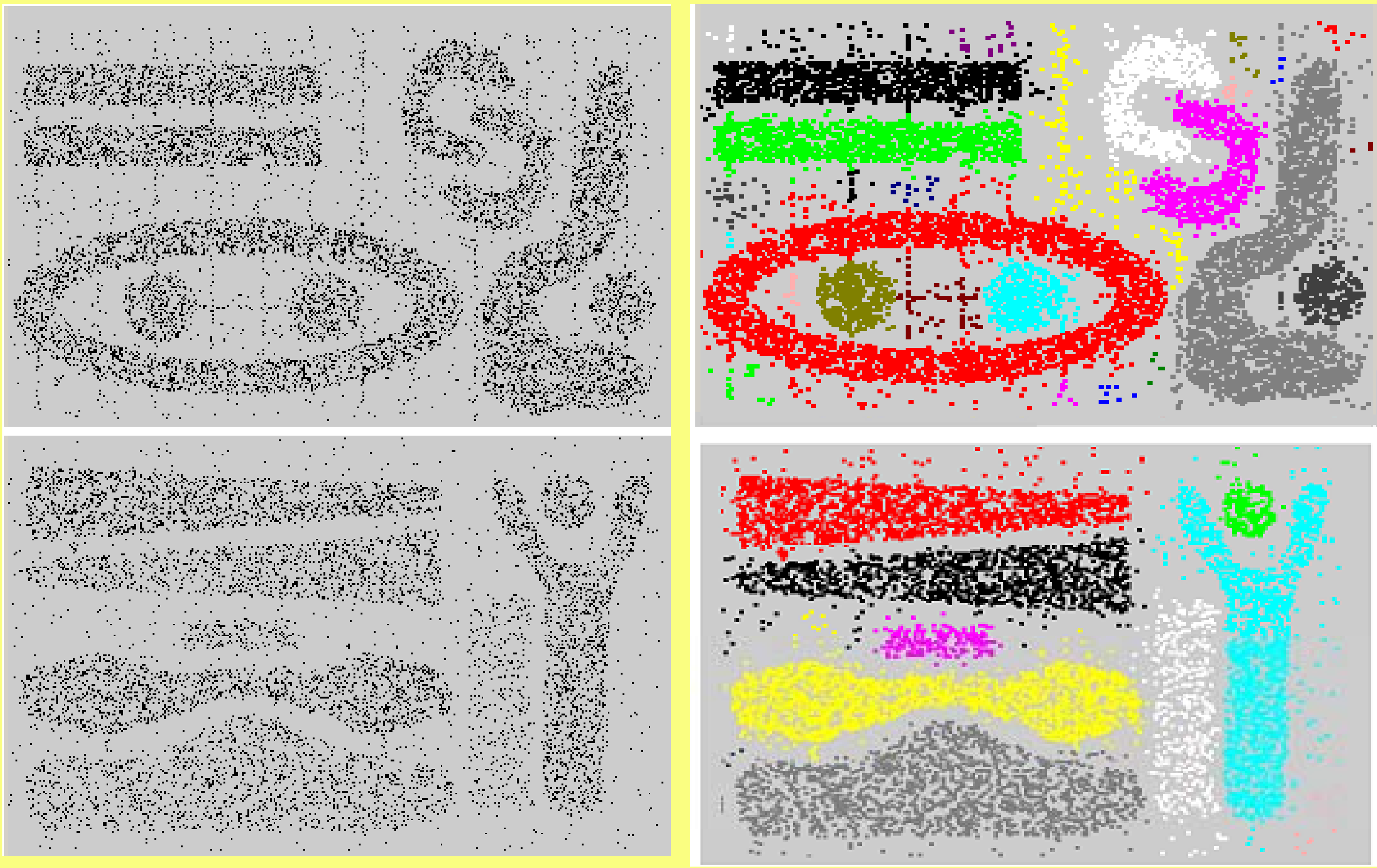
Hierarchical Grouping



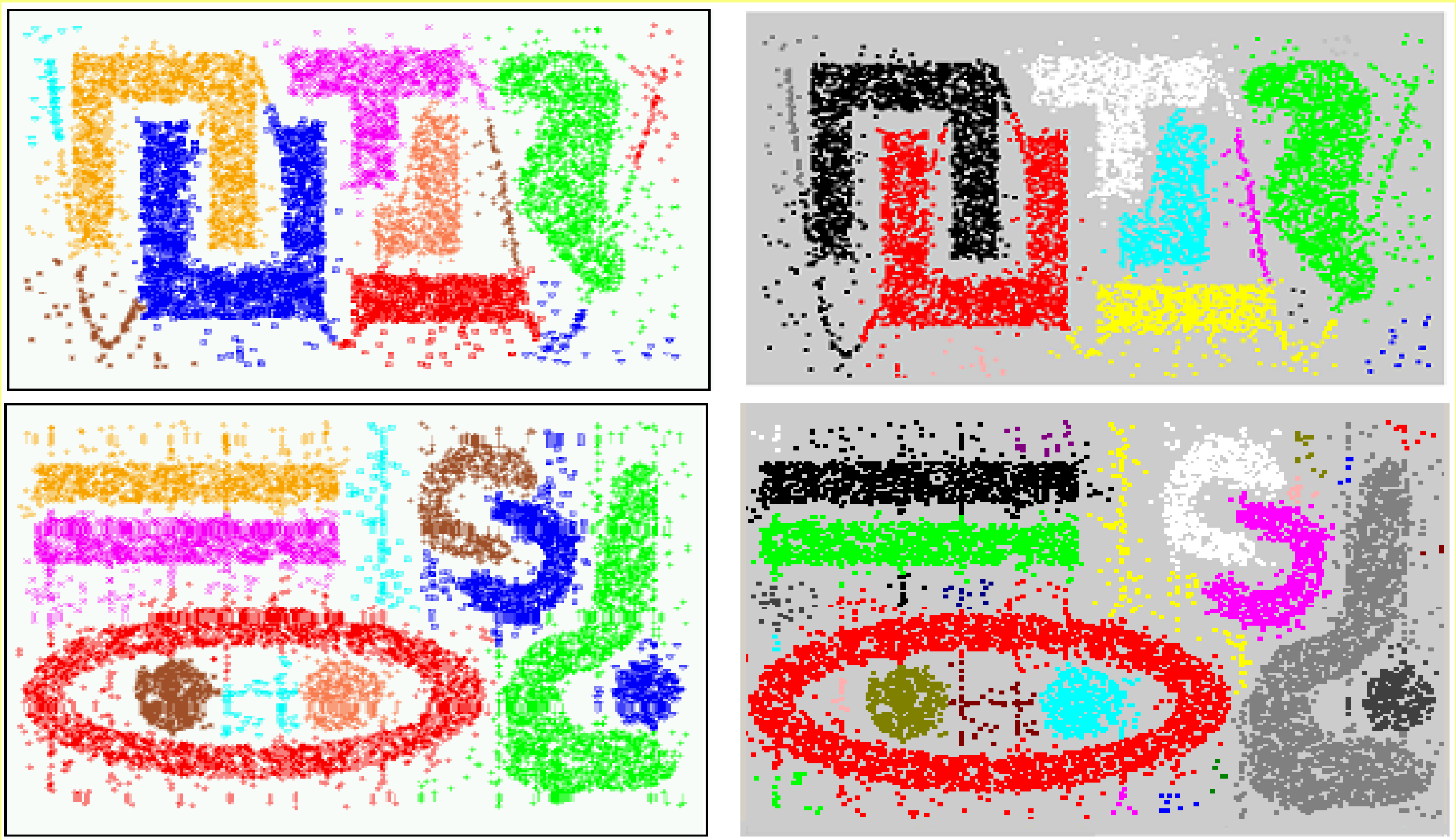
Clustering Results



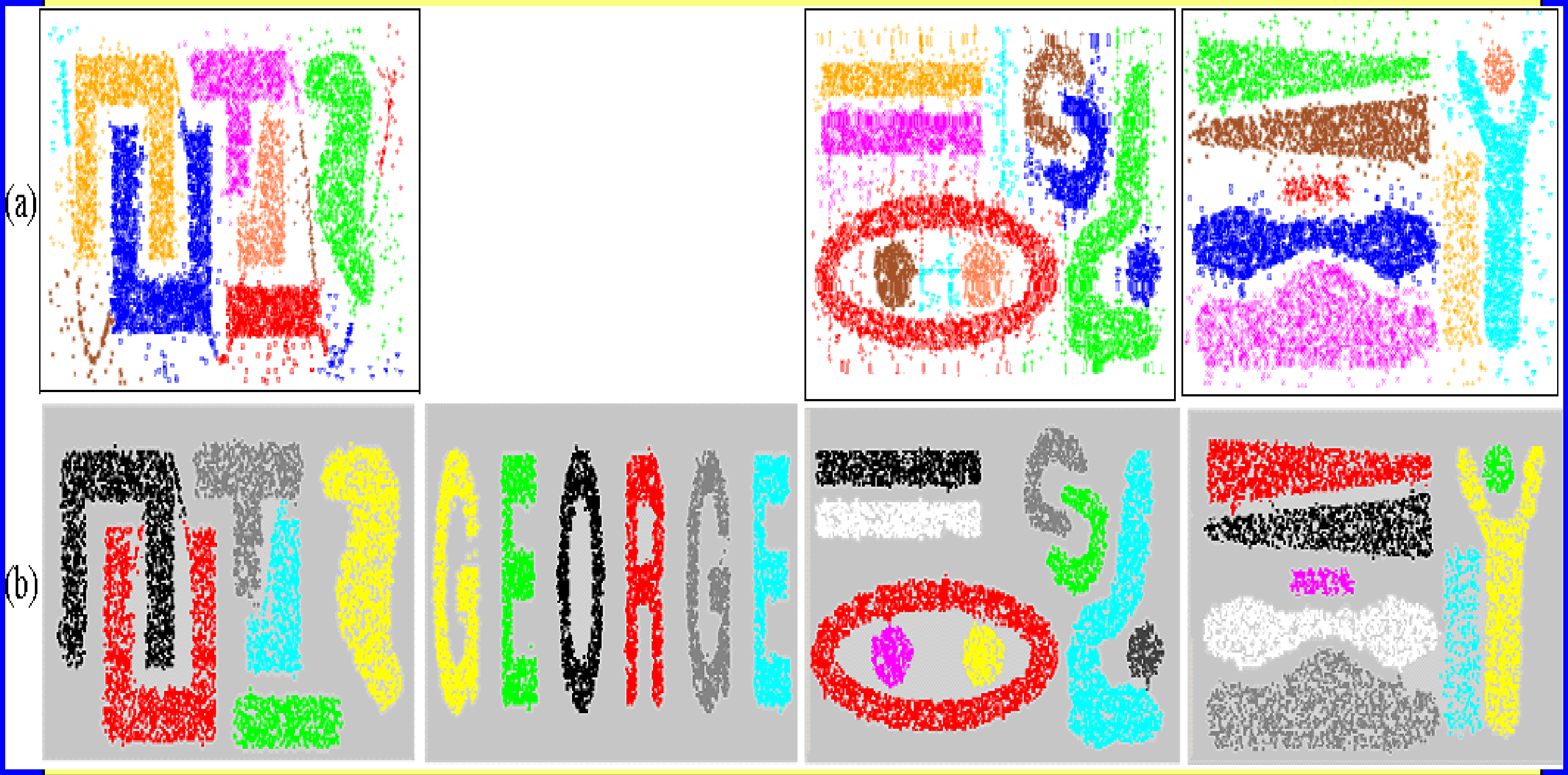
Clustering Results -- II



Compared with CAMELEON



Grouped results and comparisons



A Comprehensive Comparison

	Running Time (for *n* data points and *m* initial groups)	Finding clusters of different shapes?	Minimal input parameters		Robust to noise	
			Parameters used	How to set parameter values?	Robust?	Noise Removed?
CHAMELEON	nm+nlogn+ m*m*logm	Yes	MinSize, α, k	Fixed/Trial-and-error	Yes	No
Random Walk	nlogn	Yes	CE, NS, and weight thresholds	Fixed/Trial-and-error	Yes	Yes
SNN	n*n	Yes	k, MinPts, Eps	Fixed/Trial-and-error	Yes	Yes
CLEAN	nlogn	Yes	Km, Kc	Learned/Visualized	Yes	Yes